INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

Exploración de las preferencias de usuarios expertos sobre las regiones de importancia como explicaciones en clasificación de actividades en vídeo

Exploring expert user preferences regarding importance heatmaps as explanations in video activity classification

F. Xavier Gaya-Morey
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
francesc-xavier.gaya@uib.es

Jose M. Buades-Rubio
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
josemaria.buades@uib.es

Scott MacKenzie
Electrical Engineering and Computer Science
York University
Toronto, Ontario, Canadá
mack@yorku.ca

Raquel Lacuesta
Informática e Ingeniería de Sistemas
Universidad de Zaragoza I3A
Teruel, Aragón, España
lacuesta@unizar.es

Cristina Manresa-Yee
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
cristina.manresa@uib.es

Recibido: 16.11.2025 | Aceptado: 01.12.2025

Palabras Clave

inteligencia artificial explicable
evaluación
XAI centrado en el ser humano
métodos XAI basados en vídeo

Resumen

Aunque existen numerosos métodos de inteligencia artificial explicable (XAI), todavía hay una falta de estudios que analicen cómo los usuarios perciben la explicabilidad y la confiabilidad que estos ofrecen. Consecuentemente, es difícil determinar cuáles son los métodos de XAI más adecuados en función de las preferencias y necesidades de los usuarios. En este trabajo, usuarios expertos en IA evaluaron seis métodos XAI basados en perturbación, aplicados a través de tres redes y dos conjuntos de datos para el reconocimiento de actividades en vídeo. Para ello, se pidió a los expertos puntuar cómo de razonables fueron las explicaciones, en base a las regiones del vídeo señaladas como importantes. Los resultados muestran la preferencia por el método RISE adaptado a vídeo, mientras que identifican el método de predictores univariados adaptado a vídeo como el menos razonable. Estos hallazgos ofrecen a investigadores y profesionales una visión sobre los métodos de XAI preferidos en vídeo, al tiempo que amplían la comprensión de la explicabilidad de la IA desde una perspectiva centrada en el usuario.

Keywords

explainable artificial intelligence
evaluation
human-centered XAI
video-based XAI methods

Abstract

Many explainable artificial intelligence (XAI) methods exist; however, there is a lack of user evaluations on explainability or trustworthiness. Consequently, it remains unclear which XAI methods are appropriate based on users and their preferences. In this study, AI experts evaluated six removal-based XAI methods applied across three networks and two datasets for video-based activity recognition. For this purpose, the experts scored the reasonableness of the explanations, based on the video regions indicated as the most relevant. Experts consistently preferred the video-adapted RISE method, while identifying the video-adapted univariate predictors method as the least preferred. These findings provide insight for researchers and practitioners on the preferred XAI methods to use with videos, while also expanding the understanding of XAI methods from a human perspective.

1. Introducción

La creciente presencia de la inteligencia artificial (IA) en múltiples dominios (Dong et al., 2021), pone de relieve la importancia de garantizar su explicabilidad, de modo que las decisiones automatizadas puedan entenderse, evaluarse y, cuando proceda, cuestionarse por los usuarios. Tal como dicta el marco de trabajo HCAI (Human-Centered Artificial Intelligence framework), los métodos que identifican cuándo es necesaria la acción automática y cuándo la humana, y que evitan el exceso de peso de estas acciones son más propensas a producir diseños fiables, seguros y confiables (Shneiderman, 2020).

Desde la llegada del aprendizaje profundo (d*eep learning)*, se han desarrollado numerosos métodos de IA explicable (XAI) con el propósito de hacer inteligible la toma de decisiones de los modelos a los usuarios humanos (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). Sin embargo, pese al notable avance técnico en técnicas de explicabilidad, todavía es limitado el número de estudios que investigan cómo los usuarios finales perciben estas explicaciones, qué necesidades de explicabilidad tienen y en qué contextos las valoran positivamente.

En este sentido, resulta imprescindible no sólo diseñar nuevos métodos de XAI, sino también incorporar al usuario final desde el inicio del proceso de diseño de las explicaciones. De hecho, la literatura ha comenzado a adoptar un enfoque de IA explicable centrada en el ser humano (Human-Centred XAI, HCXAI), que integra factores humanos en la investigación y desarrollo de explicaciones de IA (Schoonderwoerd et al., 2021; Hong y Park, 2025; Ridley, 2025).

En consecuencia, avanzar en el desarrollo de HCXAI implica alinear las explicaciones de IA con los usuarios específicos, considerando sus niveles de especialización, sus tareas o los entornos de uso, y adaptar las presentaciones al contexto. Esta orientación promueve no sólo la comprensibilidad y la usabilidad, sino también la confianza, la colaboración humano-IA y, la adopción ética y eficaz de los sistemas inteligentes. Estudios recientes muestran además que los criterios para una explicación significativa no se limitan a la fidelidad técnica del modelo, sino que incluyen dimensiones como ser comprensible, accionable, concisa, coherente con la tarea del usuario y adaptada a sus expectativas (Kim et al., 2024).

Aunque las ciencias sociales han estudiado ampliamente los procesos mediante los cuales las personas generan y comprenden explicaciones, la investigación en XAI suele fundamentarse en la intuición de los propios investigadores acerca de qué constituye una "buena" explicación (Miller, 2019). Esta estrategia tiende a pasar por alto aspectos esenciales como la comprensión humana, los perfiles y necesidades de los destinatarios y los factores contextuales en torno a la explicación. Estudios recientes evidencian una baja utilización de los métodos centrados en el ser humano en el diseño de sistemas XAI (Kaplan et al., 2024; Mohseni et al., 2018; Rong et al., 2024). La literatura indica que este tipo de enfoques resultan útiles para orientar las decisiones técnicas impulsadas por las necesidades y perspectivas de los usuarios, al tiempo que permiten identificar limitaciones en los métodos existentes y proporcionar marcos conceptuales que promuevan una XAI compatible con los humanos (Liao & Varshney, 2022).

Además, a nivel de evaluación de XAI, existe una falta de marcos teóricos, metodológicos y métricas estandarizadas que permitan evaluar en qué medida los métodos de XAI ofrecen una explicabilidad útil para las personas (Floridi et al., 2018; Hoffman et al., 2019; Miró-Nicolau et al., 2024). Las escasas evaluaciones empíricas con usuarios disponibles suelen carecer de fundamentos provenientes de las ciencias cognitivas y sociales (Rong et al., 2024) y no siguen protocolos sistemáticos para medir, cuantificar y comparar la explicabilidad de los sistemas de IA (Burkart & Huber, 2021). Además, los estudios que evalúan empíricamente los métodos XAI con tareas, usuarios y contextos específicos muestran diferentes necesidades y preferencias de los usuarios (Dodge et al., 2019; Ehsan et al., 2024; Szymanski et al., 2021). En esta misma dirección, (Wells & Bednarz, 2021) llevaron a cabo una revisión sistemática examinando estudios XAI con un especial interés en el usuario, revelando que muchos estudios no involucraban a usuarios, e incluso cuando se realizaban pruebas con usuarios, a menudo se omitían detalles clave, como el número de participantes, los métodos de reclutamiento o el nivel de experiencia de los participantes en aprendizaje automático. Esto limita la transparencia y la reproducibilidad de las evaluaciones.

AL trabajar específicamente con datos visuales (es decir, imágenes o vídeos), existen evaluaciones de métodos XAI con usuarios para imágenes (Aechtner et al., 2022; Alqaraawi et al., 2020; Heimerl et al., 2020; Manresa-Yee et al., 2024), pero, hasta donde sabemos, no hay trabajos que aborden vídeos. Por lo tanto, es necesario investigar para comprender completamente el impacto de los métodos XAI para vídeo y la efectividad de las explicaciones.

El objetivo de este trabajo es realizar un estudio cuantitativo con expertos que evalúa seis métodos XAI basados en perturbación en vídeo, aplicados sobre tres redes y dos

conjuntos de datos. Para lograr esto, se adaptan seis métodos XAI, ampliamente utilizados originalmente para explicaciones locales basadas en imágenes, al dominio del vídeo. Para entender las diferencias, se generan explicaciones para tres redes con arquitecturas variadas, incluyendo *transformers* y modelos convolucionales. Además, se utilizan dos conjuntos de datos para el reconocimiento de acciones humanas disponibles públicamente: uno grabado en un entorno controlado y otro que comprende vídeos de entornos no controlados extraídos de YouTube.

Tras cuantificar las preferencias de los usuarios respecto a los seis métodos XAI, los resultados muestran acuerdo tanto en los métodos preferidos como con los menos preferidos. Estos hallazgos proporcionan, a futuros investigadores y profesionales, guías de diseño concretas, respaldadas por las elecciones de los usuarios.

El artículo se organiza de la siguiente manera: la Sección 2 proporciona una revisión de los conceptos clave relacionados. La Sección 3 detalla el sistema impulsado por IA, incluidos los conjuntos de datos, las redes neuronales y los métodos XAI utilizados. La Sección 4 describe la metodología, cubriendo los participantes, el aparato, el procedimiento y el diseño del estudio. Los resultados se presentan en la Sección 5, seguidos de una discusión en la Sección 6. Finalmente, la Sección 8 concluye y destaca posibles direcciones para futuras investigaciones.

2. Trabajo relacionado

2.1. Inteligencia artificial explicable centrada en el ser humano

XAI comprende un conjunto de métodos y técnicas orientados a mejorar la transparencia y la interpretabilidad de las decisiones y del interno de los modelos de inteligencia artificial. A medida que estos sistemas se vuelven más complejos, especialmente en el ámbito del aprendizaje profundo, comprender cómo producen sus resultados se vuelve más complejo. Las técnicas de XAI buscan reducir esta dificultad al ofrecer información sobre el comportamiento del modelo, lo que facilita que los usuarios interpreten, evalúen y confíen en las decisiones generadas por la IA (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020).

Las técnicas de explicabilidad se categorizan generalmente a lo largo de varias dimensiones clave: enfoque en datos o en modelo, explicaciones directas o *post-hoc*, alcance global o local, y presentación estática o interactiva (Arya et al., 2020). Primero, las explicaciones pueden tener como objetivo aclarar las propiedades de los datos de entrada o el comportamiento del modelo en sí. Al explicar el modelo, la distinción es entre modelos directamente interpretables (por ejemplo, regresión lineal, o árboles de decisión) y la explicabilidad *post-hoc*, que se aplica después de que se entrene el modelo. Además, las explicaciones pueden dirigirse a predicciones individuales (local) o al comportamiento del modelo en su conjunto (global). Finalmente, las explicaciones pueden ser estáticas o, como recomienda Miller (2019), diseñadas para apoyar la participación interactiva del usuario para una comprensión más profunda.

La XAI centrada en el ser humano se basa no solo en la explicabilidad técnica, sino también en alinear las explicaciones con las necesidades humanas, los procesos cognitivos y el contexto (Barda et al., 2020; Liao et al., 2020). En lugar de asumir que una explicación es suficiente, la XAI centrada en el ser humano enfatiza la usabilidad, la interpretabilidad y la relevancia para diversos usuarios, incluidos los no expertos (Lopes et al., 2022). El objetivo es crear explicaciones que sean intuitivas y conscientes del contexto y, por lo tanto, apoyen la toma de decisiones para mejorar la colaboración entre humanos y sistemas de IA (Ehsan et al., 2022; Liao & Varshney, 2022). Esta perspectiva reconoce que la efectividad de una explicación depende tanto del usuario como del método en sí.

2.2. XAI aplicado a datos de vídeo

Mientras que los métodos XAI basados en imágenes se han estudiado exhaustivamente en la literatura (por ejemplo, Lundberg & Lee, 2017; Petsiuk et al., 2018; Ribeiro et al., 2016), los métodos XAI basados en vídeo, particularmente los agnósticos al modelo, permanecen relativamente inexplorados, en parte debido a los desafíos únicos que plantea el vídeo, como la mayor dimensionalidad. Sin embargo, los métodos agnósticos del modelo son valiosos porque ofrecen flexibilidad y amplia aplicabilidad para escenarios del mundo real.

Los enfoques agnósticos más empleados se basan en la eliminación o modificación de la entrada (*removal-based explanations*). Su objetivo común es estimar la relevancia de las características de entrada analizando los cambios en las predicciones cuando se perturban o eliminan partes de la entrada. Dado que las modificaciones se realizan directamente sobre el espacio de entrada, estos métodos no dependen del modelo, y la importancia de cada característica se infiere únicamente a partir de la variación en la salida. En el caso de los datos de vídeo, la elevada dimensionalidad y la naturaleza espacio-temporal del contenido obligan a introducir adaptaciones, como el uso de grupos de píxeles o

regiones (superpíxeles) y la extensión del análisis a lo largo del tiempo.

Entre los métodos agnósticos más representativos (y los empleados en nuestros experimentos) se encuentran LIME, SHAP, RISE, LOCO, *univariate predictors* y *occlusion sensitivity*.

LIME (*Local Interpretable Model-agnostic Explanations*) explica las predicciones de un clasificador mediante un modelo interpretable local, como una regresión lineal o un árbol de decisión. Particiona la imagen (o fotograma) en regiones llamadas superpíxeles y genera muestras perturbadas ocluyendo aleatoriamente parte de ellas. La relevancia de cada región se estima en función del cambio observado en la confianza de la predicción (Ribeiro et al., 2016).

SHAP (*SHapley Additive exPlanations*) introduce los valores de Shapley como medida unificada de importancia de las características. Estos valores representan el cambio esperado en la predicción del modelo al considerar o excluir una característica. Kernel SHAP, una variante eficiente y aproximada, combina elementos de LIME para garantizar exactitud local y coherencia en la asignación de relevancias (Lundberg y Lee, 2017).

RISE (*Randomized Input Sampling for Explanation*) genera máscaras binarias aleatorias sobre una cuadrícula de baja resolución, que posteriormente se interpola hasta las dimensiones del fotograma. Cada muestra se obtiene ocluyendo diferentes regiones y midiendo el efecto en la probabilidad de clase. La relevancia final se calcula promediando las predicciones en aquellas muestras donde las regiones permanecen visibles, produciendo un mapa de calor de relevancia positiva (Petsiuk et al., 2018).

LOCO (*Leave-One-Covariate-Out*) evalúa la influencia de cada característica eliminándola del conjunto de entrenamiento y observando la variación resultante en las predicciones. A diferencia de los anteriores, proporciona explicaciones globales del modelo en lugar de locales (Lei et al., 2018).

Los *univariate predictors* proponen evaluar el impacto de cada variable de manera independiente, optimizando la interpretabilidad y reduciendo la complejidad computacional. Como LOCO, este enfoque ofrece explicaciones globales (Guyon & Elisseeff, 2003).

Occlusion sensitivity calcula la importancia de los píxeles desplazando un parche (por ejemplo, gris) sobre la imagen y midiendo el cambio en la confianza de la predicción. En el contexto del vídeo, este método se extiende añadiendo una dimensión temporal al kernel usado para la oclusión, lo que permite capturar la relevancia espaciotemporal y adaptarse mejor a secuencias con cortes, movimientos de cámara u objetos que entran o salen del campo de visión (Zeiler & Fergus, 2014).

Estos métodos, con las adaptaciones pertinentes para ser extrapolados a su uso en vídeo (Gaya-Morey et al., 2024) constituyen las principales aproximaciones agnósticas utilizadas actualmente para proporcionar explicaciones interpretables en tareas de reconocimiento basadas en vídeo.

2.3. Estudios de usuario en datos visuales

La mayoría del trabajo existente en evaluación de métodos XAI se centra en métricas automáticas, a menudo pasando por alto cómo los usuarios reales interpretan, confían o se benefician de estas explicaciones (Miller, 2019). En consecuencia, relativamente pocos estudios de usuario evalúan explicaciones de imágenes y vídeos con participantes humanos. Esta brecha es especialmente pronunciada en el dominio del vídeo, donde la dimensión temporal añade complejidad a la interpretación humana. Evaluar las explicaciones con usuarios reales es crucial para comprender su utilidad práctica, mejorar las opciones de diseño y garantizar que dichos sistemas se alineen con el razonamiento y la toma de decisiones humanos.

Con respecto a los estudios de usuario sobre métodos XAI aplicados a imágenes, Aechtner et al. (2022) estudiaron la percepción de los usuarios sobre las explicaciones locales frente a las globales, mostrando la preferencia por las explicaciones locales de los usuarios poco habituados al uso de IA. Manresa-Yee et al. (2024) también estudiaron las explicaciones locales frente a las globales, involucrando a 104 usuarios en un estudio que analizaba aspectos como la confianza percibida o la comprensión. Se observaron puntuaciones más altas para las combinaciones de ambas explicaciones. Alqaraawi et al. (2020) investigaron el rendimiento de los mapas de saliencia en un estudio con usuarios, mostrando una preferencia por LRP, y señalaron la ayuda limitada de las explicaciones para predecir la salida de la red para nuevas imágenes o para identificar las características de la imagen a las que el sistema es sensible.

Selvaraju et al. (2017) exploraron si las explicaciones de Grad-CAM ayudaron a los usuarios a establecer una confianza adecuada en las predicciones. Sus resultados mostraron que Grad-CAM permitió a los usuarios no entrenados diferenciar con éxito una red profunda "más fuerte" de una "más débil", incluso cuando producían predicciones idénticas.

En nuestra revisión de la literatura para XAI aplicado a vídeo, no se encontró ningún trabajo que evalúe o compare diferentes explicaciones en vídeos desde una perspectiva humana.

3. Sistema de reconocimiento de actividades y métodos XAI

Para evaluar las preferencias del usuario por los métodos XAI de vídeo, se creó un conjunto de que combina tres redes, dos conjuntos de datos y seis métodos XAI. Esto permitió introducir variaciones en las explicaciones, para identificar la influencia de estos tres factores.

3.1. Conjuntos de datos

Se seleccionaron dos conjuntos de datos con características distintas para entrenar los modelos y evaluar los métodos XAI: Kinetics 400 (Kay et al., 2017) y EtriActivity3D (Jang et al., 2020).

El conjunto de datos Kinetics 400 es una colección a gran escala de vídeos de YouTube que cubre 400 categorías, con al menos 400 videoclips por clase. El conjunto de datos se centra en diversas actividades humanas, incluidas las interacciones entre personas y las interacciones con objetos. Presenta una amplia variedad de participantes, entornos y objetos, junto con desafíos como movimientos de cámara y cortes dentro del mismo clip, lo que contribuye a su complejidad.

En contraste, EtriActivity3D es un conjunto de datos más especializado que contiene 112.620 vídeos, clasificados en 55 actividades. Se centra en tareas cotidianas realizadas por 100 individuos, la mitad de los cuales tienen más de 64 años, lo que permite el estudio sobre demografía de edad avanzada. Los vídeos se capturaron en entornos domésticos, incluyendo múltiples habitaciones, y desde ocho cámaras fijas, asegurando una grabación estable y sin cortes para cada clip. Se trata, por tanto, de una configuración muy controlada que permite obtener un conjunto de datos menos complejo que Kinetics 400.

3.2. Redes neuronales

Se utilizaron tres redes: TimeSformer (Bertasius et al., 2021), TANet (Liu et al., 2021) y TPN (Yang et al., 2020). Estas redes son representantes de diferentes arquitecturas neuronales, como los *transformers* y las redes convolucionales, lo que nos permite explorar si dicha arquitectura tiene un impacto en las preferencias de los usuarios. La elección de las redes se justifica por su rendimiento en tareas de clasificación de acciones y su disponibilidad pública dentro de la caja de herramientas de código abierto basada en PyTorch MMAction2 para el análisis de vídeo (OpenMMLab, 2020).

TimeSformer, una variante del *Vision Transformer*, captura características espaciotemporales procesando parches a nivel de fotograma. TANet incorpora un Módulo Adaptativo Temporal (TAM) dentro de su marco CNN 2D, lo que permite la captura de dinámicas temporales tanto a corto como a largo plazo utilizando un mecanismo adaptativo de dos niveles. La Red Piramidal Temporal (TPN), por otro lado, extrae e integra información espacial, temporal y semántica utilizando un reescalado jerárquico, mejorando el rendimiento para tareas con variabilidad temporal. Para TANet y TPN, utilizamos la arquitectura ResNet50 como *backbone*.

Para el conjunto de datos Kinetics 400, utilizamos pesos pre-entrenados disponibles en MMAction2. Para el conjunto de datos EtriActivity3D, realizamos *fine tuning* de las redes utilizando los pesos pre-entrenados de Kinetics 400, entrenando durante 10 épocas con validación cruzada con k=5.

3.3. Métodos XAI

Dado que adoptamos redes con arquitecturas variables, optamos por métodos XAI agnósticos del modelo, que generan explicaciones independientemente del modelo subyacente. Específicamente, empleamos una versión adaptada a vídeo de métodos XAI agnósticos del modelo ampliamente utilizados, que se encuentra disponible públicamente[1]. Dichos métodos incluyen adaptaciones de LIME (Ribeiro et al., 2016) (Video LIME), Kernel-SHAP (Lundberg & Lee, 2017) (Video Kernel-SHAP), RISE (Petsiuk

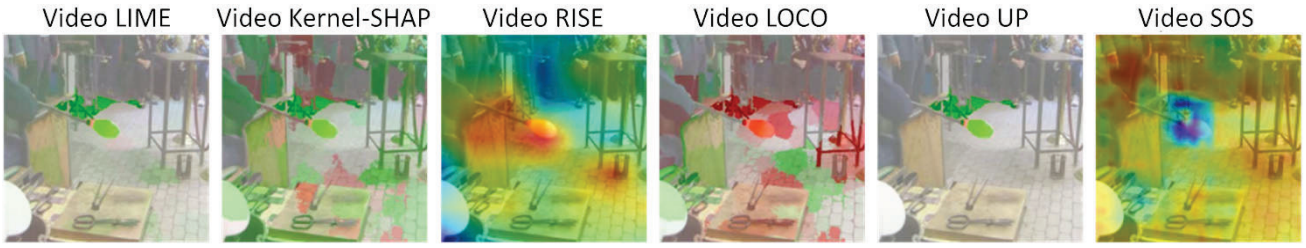

Figura 1: Ejemplo de explicaciones calculadas usando el modelo TimeSformer y el conjunto de datos Kinetics 400, utilizando los diferentes métodos. Solo se muestra un fotograma del momento de máxima relevancia por método.

et al., 2018) (Video RISE), *Occlusion sensitivity* (Zeiler & Fergus, 2014) (Video SOS), LOCO (Lei et al., 2018) (Video LOCO) y predictores univariados (Guyon & Elisseeff, 2003) (Video UP). Relacionándolos con las dimensiones de XAI mencionadas anteriormente, estos métodos tienen como objetivo explicar el modelo y se caracterizan por ser *post-hoc*, de alcance local y estáticos en la presentación (Gaya-Morey et al., 2024).

La operación de estos métodos implica cuatro pasos principales: (1) segmentar el vídeo de entrada en regiones que consisten en píxeles de diferentes fotogramas, (2) ocluir estas regiones y pasar el vídeo modificado a través del modelo, donde las predicciones cambian según las regiones ocluidas, (3) resumir la relevancia de cada región para la predicción objetivo y (4) visualizar las explicaciones. Los parámetros exactos utilizados en cada paso dependen del método. La aplicación de estos métodos XAI para explicar la predicción de un modelo para un vídeo dado produce una explicación en forma de vídeo, dentro del cual cada píxel representa la relevancia del píxel correspondiente en el vídeo original. La Figura 1 muestra un ejemplo de explicación utilizando cada método.

3.4. Explicaciones de vídeos

Para la evaluación, se seleccionó una muestra aleatoria de 30 vídeos de cada conjunto de datos: Kinetics 400 y EtriActivity3D. Por consistencia, solo se incluyeron vídeos que fueron clasificados correctamente por las tres redes utilizadas en el estudio. Si un vídeo fue mal clasificado por alguna red, fue reemplazado por otro vídeo seleccionado al azar. Para asegurar una comparación justa, se impusieron condiciones iguales para los métodos en la medida de lo posible, como el número de características, muestras y tipo de oclusión.

Cada uno de los 30 vídeos de ambos conjuntos de datos se procesó a través de las tres redes: TimeSformer, TPN y TANet. Para cada predicción, se generaron explicaciones utilizando los seis métodos XAI descritos anteriormente. Esto resultó en 6 participantes × 6 métodos XAI × 3 redes × 2 conjuntos de datos × 30 vídeos por condición =1.080 explicaciones a lo largo del experimento.

Para mejorar la claridad y la interpretabilidad de las explicaciones, solo se retuvo el 30% superior de las regiones más relevantes, filtrando las áreas menos significativas (puede apreciarse más adelante en las Figuras Figura 2 y Figura 5). Además, se aplica estiramiento de histograma para asegurar que las explicaciones utilizaran todo el rango del espectro de color, haciendo que las visualizaciones fueran más claras. Además, se eliminaron los valores de relevancia negativos por dos razones principales: simplificar la información presentada a los usuarios que evalúan las explicaciones y estandarizar las salidas en todos los métodos XAI, ya que no todos los métodos proporcionan puntuaciones de relevancia tanto positivas como negativas.

4. Método

Se presentaron explicaciones a los usuarios para evaluar sus preferencias.

4.1. Participantes

Seis voluntarios (tres mujeres) del entorno universitario participaron en el estudio. Las edades oscilaron entre 24 y 47 años (media = 34,7; DT = 10,4). Dichos participantes tienen una amplia experiencia tanto en IA como en XAI, con su conocimiento basado en años de investigación especializada y aplicaciones prácticas. Dos de los expertos, los más jóvenes, trabajaron en IA durante al menos tres o cuatro años y han pasado los últimos dos años trabajando en XAI. Los expertos más experimentados tienen una amplia trayectoria tanto en IA como en XAI, habiendo trabajado en esta última área durante un mínimo de cinco años. Además, tres de los expertos centran su investigación en Interacción Persona-Ordenador (IPO). Su investigación abarca un amplio espectro, incluida la visión por computador y el aprendizaje profundo aplicados a problemas de (IPO). Aunque todos los participantes tenían experiencia con IA y XAI, su familiaridad no se extendía a todos los métodos XAI.

4.2. Aparato

Se desarrolló una interfaz de usuario para mostrar el vídeo, su clase asociada, la explicación y un mapa de color correspondiente para ayudar a los usuarios en su evaluación (ver Figura 2). La interfaz muestra las 1.080 explicaciones, cada visualización muestra solo una explicación.

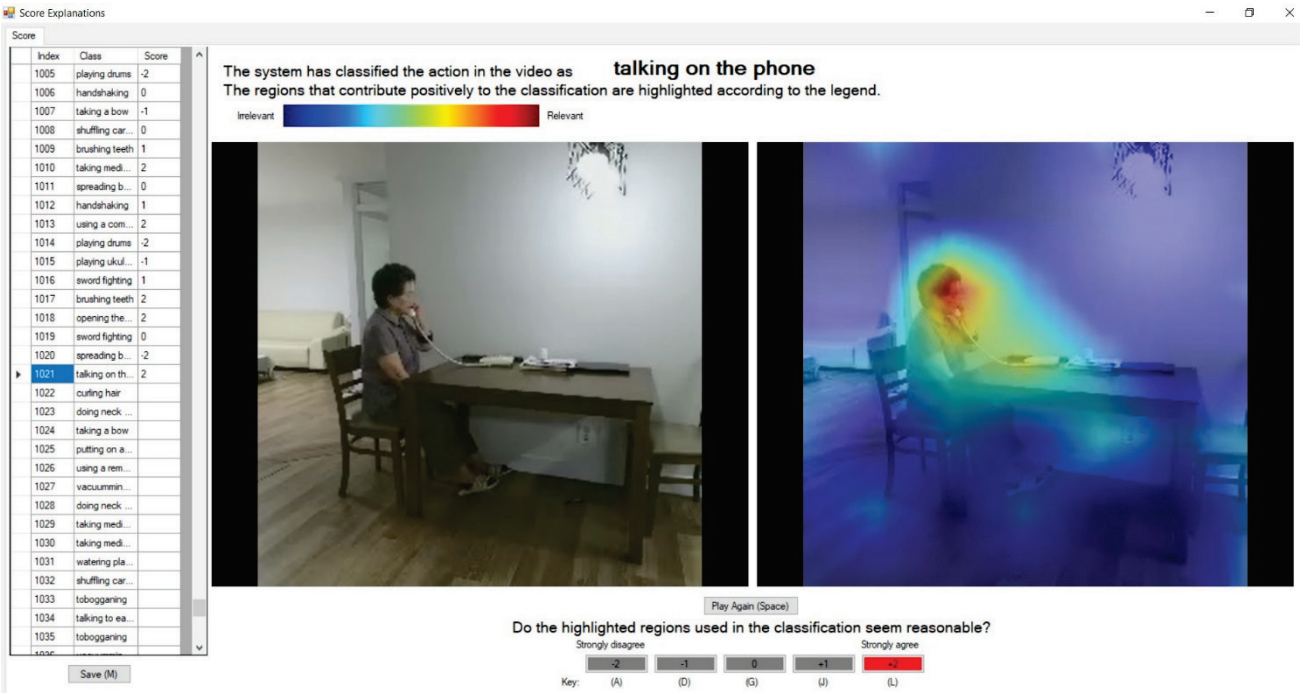

Figura 2: Interfaz de usuario para puntuar las explicaciones generadas por el método XAI. A la izquierda, una lista desplazable de vídeos con su clase correspondiente y las puntuaciones asignadas por el usuario. En la parte superior, se muestra información pertinente sobre la explicación actual, incluida la clase y el mapa de color que representa la explicación. En el centro, se presenta el vídeo que se está explicando (izquierda) junto con su explicación correspondiente (derecha). En la parte inferior, se presenta al usuario una pregunta y opciones de respuesta, con la respuesta seleccionada en rojo.

La pregunta planteada a los usuarios durante la evaluación fue: "¿Las regiones resaltadas utilizadas en la clasificación parecen razonables?" Las opciones de respuesta se presentan en una escala de Likert con valores de entre -2 (totalmente en desacuerdo) y +2 (totalmente de acuerdo). Por lo tanto, las puntuaciones positivas resaltan el alineamiento entre las regiones importantes según el método y el usuario, y las negativas, el no alineamiento. La pregunta busca determinar si las regiones resaltadas se alinean con las percepciones de los usuarios al identificar acciones específicas en el vídeo. Para mitigar cualquier posible sesgo, las explicaciones se presentan en orden aleatorio y sin información sobre la red, el conjunto de datos o el método XAI. Esto asegura una evaluación "ciega".

4.3. Procedimiento

El estudio se llevó a cabo utilizando un portátil con el programa instalado localmente, que se permitió a los participantes llevar a casa. A cada participante se le encomendó la tarea de evaluar 1.080 explicaciones, un proceso que requirió a cada usuario aproximadamente de 3 a 4 horas. Para acomodar esto, se les dio a los participantes la flexibilidad de pausar y reanudar la evaluación a su conveniencia.

Las explicaciones se presentaron a todos los participantes en el mismo orden, con la siguiente explicación mostrada automáticamente después de evaluar la anterior. Sin embargo, los participantes tuvieron la flexibilidad de poder navegar libremente entre explicaciones, lo que les permitió volver a visitar, reevaluar y actualizar sus puntuaciones según fuera necesario.

Para cada método, se calculó la puntuación media para todos los participantes para cada método XAI. Además, creamos gráficos de barras agregados de las puntuaciones de los participantes por método, conjunto de datos y red y analizamos la significancia estadística de los diferentes factores.

4.4. Diseño

El estudio siguió un diseño intrasujetos de 6×3×2 con las siguientes variables independientes y niveles:

- Método XAI (Video RISE, Video Kernel-SHAP, Video LOCO, Video LIME, Video SOS, Video UP)
- Red (TimeSformer, TANet, TPN)
- Conjunto de datos (EtriActivity3D, Kinetics400)

La variable dependiente fue la puntuación de razonabilidad en una escala Likert de 5 puntos de -2 (*totalmente en desacuerdo*) a 2 (*totalmente de acuerdo*).

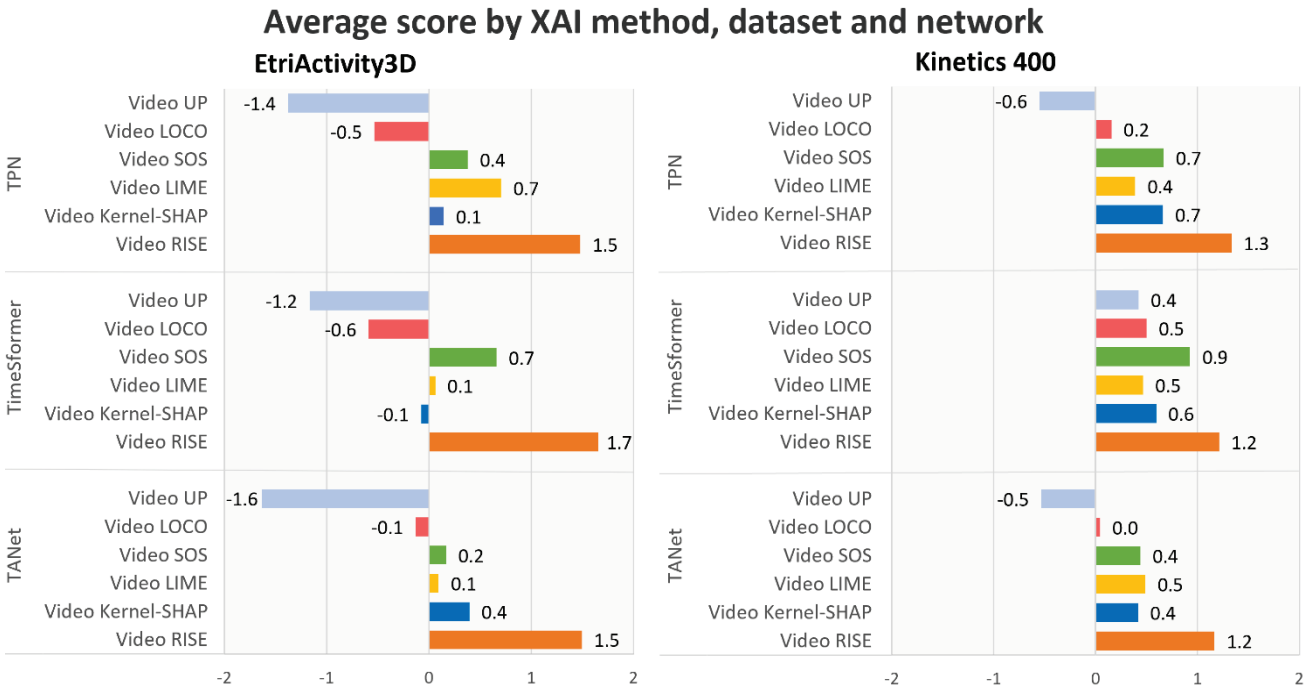

Figura 3: Promedio de puntuaciones de los usuarios agrupadas por método XAI, conjunto de datos y red.

El número total de ensayos fue de 6.480 (= 6 participantes × 6 métodos XAI × 3 redes × 2 conjuntos de datos × 30 vídeos por condición).

5. Resultados

En esta sección se presentan los resultados de las evaluaciones realizadas por parte de los participantes. La media general sobre las 6.480 explicaciones fue de 0,292. Respecto a la pregunta de interés, las respuestas muestran valores entre 0 (neutral) y 1 (ligeramente de acuerdo), por lo tanto, hubo una tendencia general de los participantes a sentir que las explicaciones de los métodos XAI se inclinaban hacia "razonables". Por método XAI, las medias, de menor a mayor, fueron de -0,806 (Video UP), -0,093 (Video LOCO), 0,368 (Video SOS), 0,356 (Video Kernel-SHAP), 0,540 (Video LIME) y 1,390 (Video RISE). Por red, las medias fueron 0,200 (TANet), 0,389 (TimeSformer) y 0,288 (TPN). Por conjunto de datos, las medias fueron 0,096 (EtriActivity3D) y 0,489 (Kinetics 400). A continuación, se describen múltiples análisis realizados por diferentes combinaciones de estas condiciones. Los resultados de la evaluación de los usuarios se presentan en la Figura 3 y la Figura 4. La Figura 3 ilustra las puntuaciones promedio por método, mientras que la Figura 4 agrega las puntuaciones por conjunto de datos, red y método.

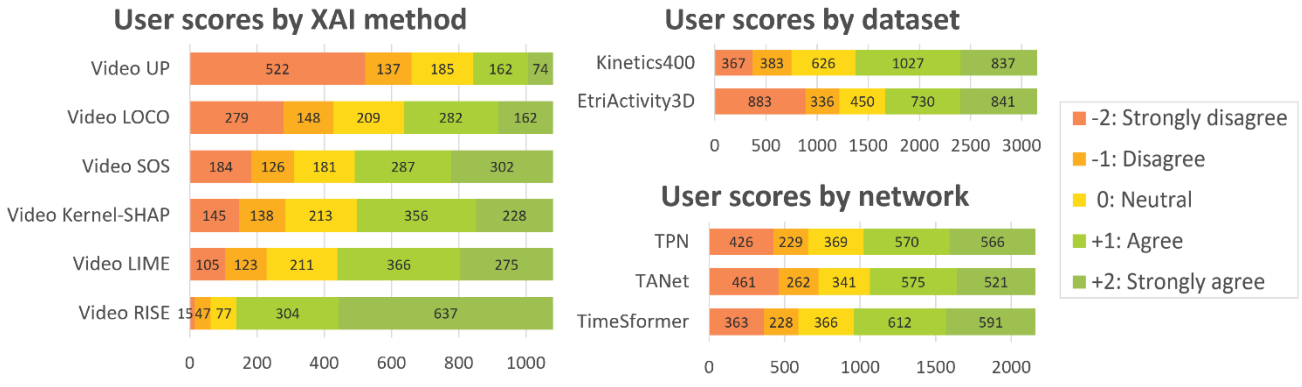

Figura 4: Recuento de puntuaciones de usuario agrupadas por método XAI, por conjunto de datos y por red.

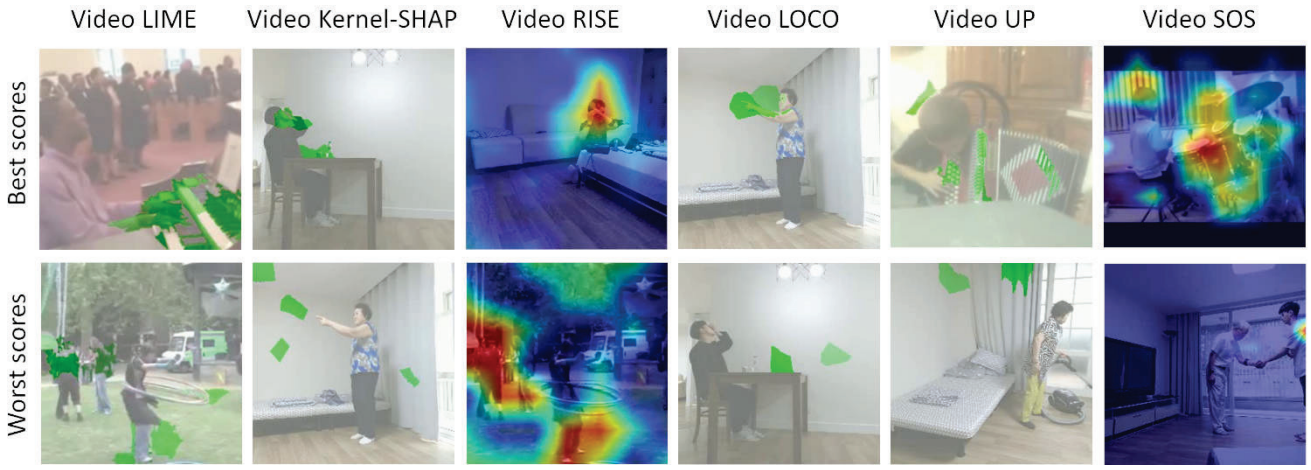

Figura 5: Imágenes de ejemplo que recibieron las puntuaciones más altas y bajas de los seis expertos. Las columnas representan diferentes métodos XAI, mientras que las filas muestran las imágenes con la puntuación más alta (arriba) y las imágenes con la puntuación más baja (abajo).

Primero, se realizó un ANOVA de tres vías para evaluar los efectos del conjunto de datos, la arquitectura de la red neuronal y el método XAI en las valoraciones de los usuarios. Se encontraron efectos principales significativos para el conjunto de datos ($F_{1,6444}$=162,83), la red ($F_{2,6444}$=12,49) y el método XAI ($F_{5,6444}$=369,43). Además, se observaron interacciones significativas entre el conjunto de datos y la red ($F_{2,6444}$=10,98), el conjunto de datos y el método XAI ($F_{5,6444}$ = 43,17), la red y el método XAI ($F_{10,6444}$=9,06), y la interacción de tres vías entre el conjunto de datos, la red y el método XAI ($F_{10,6444}$=5,61). En todos los casos, p<0,001. Estos resultados indican que la percepción del usuario no solo depende de factores individuales, sino también de sus combinaciones, con el método XAI mostrando el efecto más fuerte en las valoraciones.

Para evaluar el poder explicativo de diferentes factores, calculamos R^2 para tres modelos: uno que incluye todos los factores (conjunto de datos, red y método XAI), uno que considera solo el método XAI y uno que incluye solo el conjunto de datos y la red. El modelo completo alcanzó R^2 = 0,273, lo que indica que los factores juntos explican el 27,3% de la varianza en las valoraciones de los usuarios. El modelo que considera solo el método XAI arrojó R^2 = 0,208, confirmando que el método XAI es el factor más influyente. Por el contrario, el modelo que incluye solo el conjunto de datos y la red arrojó solo R^2 = 0,024, lo que sugiere que estos factores por sí solos contribuyen mínimamente a explicar las valoraciones de los usuarios. Además, el modelo completo tuvo el AIC (21.216) y el BIC (21.460) más bajos, lo que indica el mejor equilibrio entre la bondad de ajuste y la complejidad del modelo.

A continuación, se realizó una prueba *post hoc* de Tukey HSD para analizar las diferencias por pares entre los métodos XAI. Los resultados mostraron diferencias significativas en la mayoría de las comparaciones (p<0,05), excepto entre Video Kernel-SHAP y Video SOS (p=0,999), donde no se encontró ninguna diferencia significativa. El método Video RISE obtuvo consistentemente valoraciones significativamente más altas en comparación con otros métodos, con las mayores diferencias observadas frente a Video UP (diferencia media=2,20; p<0,001). Por el contrario, Video UP recibió valoraciones significativamente más bajas que todos los demás métodos. Estos hallazgos confirman que la elección del método XAI influye fuertemente en las valoraciones de los usuarios. Video RISE exhibe los resultados más favorables, alcanzando una puntuación promedio de 1,39 dentro del rango [-2, 2]. En segundo lugar, se encuentra Video LIME (0,54), seguido de cerca por Video SOS (0,37) y Video Kernel SHAP (0,36). Por el contrario, Video LOCO puntúa mal (-0,09), y Video UP recibe una puntuación de -0,81, la puntuación más baja.

También se realizó una prueba *post hoc* de Tukey HSD para analizar las diferencias por pares entre las tres arquitecturas de redes neuronales. Los resultados revelaron que TimeSformer recibió valoraciones significativamente más altas que TANet (diferencia media=0,19; p<0,001). Sin embargo, las diferencias entre TPN y TANet (diferencia media=0,09; p=0,119) y entre TimeSformer y TPN (diferencia media = 0,10; p = 0,056) no fueron estadísticamente significativas.

6. Discusión

La preferencia por Video RISE por parte de los expertos sugiere que colocar regiones importantes sobre la persona que realiza la acción tiene sentido para los usuarios (ver Figura 5, primera fila, tercera columna, explicación para la clase "cepillarse el pelo"). Además, las explicaciones suaves mostradas por Video RISE, sin bordes duros, fueron mejor

valoradas sobre otros métodos. Esta observación plantea la cuestión de si introducir suavidad en otros métodos, como a través de un filtro gaussiano, influiría positivamente en la calidad de la explicación según los usuarios. Si bien Video RISE logró consistentemente resultados superiores en todos los conjuntos de datos y redes, el rendimiento de otros métodos varió dependiendo de estos factores. Por ejemplo, Video UP puntuó aproximadamente un punto más alto en Kinetics 400 que en EtriActivity3D, y Video SOS funcionó mejor en TimeSformer que en otras redes. Esto sugiere que ciertos métodos XAI pueden ser más adecuados para redes neuronales o características de datos específicas.

El conjunto de datos también influyó en las valoraciones de los usuarios. En promedio, las puntuaciones para Kinetics 400 fueron 0,39 puntos más altas que las de EtriActivity3D, con el ANOVA confirmando esta diferencia como significativa ($F_{1,6444}$=162,8; p<0,001). Atribuimos la diferencia a la complejidad del conjunto de datos: Kinetics 400 presenta vídeos más complejos con movimientos de cámara, cortes y una gama más amplia de clases de acción, lo que dificulta la generación de explicaciones. En contraste, EtriActivity3D ofrece un contexto más simple para identificar regiones importantes para la clasificación, lo que probablemente influyó en las puntuaciones de los usuarios.

Con respecto a la selección de la red, observamos diferencias significativas en las puntuaciones promedio de los usuarios entre dos modelos: TimeSformer (puntuación promedio=0,39) y TANet (puntuación promedio=0,20). Sin embargo, no se encontró ninguna diferencia significativa entre TPN (puntuación promedio=0,29) y cualquiera de los otros dos. Esto demuestra que, incluso cuando se entrena en condiciones idénticas, las diferencias de arquitectura entre modelos impactan las evaluaciones de los usuarios. Por ejemplo, las explicaciones con Video UP y Video SOS recibieron consistentemente puntuaciones más altas cuando se generaron para TimeSformer en comparación con las otras dos redes, como se muestra en la Figura 3. En consecuencia, para garantizar una evaluación justa de los métodos XAI, los experimentos deben incluir múltiples redes que representen diversas arquitecturas.

7. Limitaciones del estudio

Una limitación de este estudio es el tamaño de la muestra de participantes. Sin embargo, el acuerdo unánime entre los participantes tanto en las mejores como en las peores explicaciones fortalece nuestra confianza en los hallazgos. La Figura 5 presenta ejemplos de explicaciones que recibieron por unanimidad las puntuaciones más altas y bajas.

Otra posible amenaza a la validez del estudio es la muestra de métodos XAI, redes, datasets y actividades elegidas. Aunque se ha incluido relativa variedad a cada uno de estos aspectos, al haberse constatado que algunos de ellos tienen un impacto directo en las puntuaciones del usuario (por ejemplo el dataset y la red), cabe plantearse la necesidad de incluir mayor variedad en futuros estudios, e incluso explorar qué características concretas están influyendo en las puntuaciones de los usuarios.

Finalmente, hay que mencionar que el enfoque de este estudio es sobre usuarios expertos en IA, capaces de entender explicaciones en forma de mapas de calor y sus implicaciones para con las redes neuronales usadas. Por tanto, se ha dejado para trabajo futuro la comprobación del impacto que pueda tener este factor de conocimiento del ámbito sobre las puntuaciones.

8. Conclusión

Aunque existen numerosos métodos XAI para generar explicaciones, seleccionar el método más adecuado sigue siendo un reto tanto para investigadores como para profesionales. Este estudio marca un paso adelante en la comprensión de cómo los usuarios perciben seis métodos XAI conocidos (incluidos LIME, SHAP o RISE) cuando se adaptan al dominio del vídeo. Al aplicar explicaciones en diversos conjuntos de datos y redes, se analiza la influencia de esos factores. Sorprendentemente, y aunque la muestra de expertos es pequeña, hubo consenso: Video RISE fue el preferido por los participantes, mientras que Video UP recibió las puntuaciones más bajas.

Los estudios de usuario para evaluar los métodos XAI son esenciales para obtener información sobre cómo los usuarios interactúan e interpretan las explicaciones de los sistemas de IA. Este conocimiento puede guiar las decisiones técnicas basadas en las preferencias de explicabilidad de los usuarios y ayudar a elegir un método XAI apropiado para aplicaciones del mundo real.

Sin embargo, los estudios consumen mucho tiempo y son costosos, ya que requieren gran cantidad de recursos para recopilar datos significativos. Para acelerar el proceso de evaluación, las métricas automáticas como el área bajo la curva (AUC) pueden ofrecer formas más eficientes de evaluar los métodos XAI. Sin embargo, persiste un debate sobre si el rendimiento de los métodos XAI a través de métricas objetivas debe tener prioridad sobre las preferencias del usuario al determinar su efectividad o aplicación. No obstante, es fundamental probar qué tan bien se alinean las métricas automáticas con la perspectiva del usuario. Cerrar

esta brecha asegurará que el proceso de evaluación siga siendo efectivo y representativo de las experiencias reales del usuario.

El trabajo futuro implicará evaluaciones con una muestra más grande de usuarios para validar y probar aún más nuestros resultados. Además, incorporar una mayor diversidad de participantes, como variaciones en el conocimiento de IA, la edad y otros datos demográficos, proporcionará una comprensión más profunda de los métodos XAI desde una perspectiva humana.

Agradecimientos

Este trabajo forma parte del proyecto PID2023-149079OB-I00 (EXPLAINME), financiado por MICIU/AEI/10.13039/501100011033/ y FEDER (UE), y del proyecto PID2022-136779OB-C32 (PLEISAR), financiado por MICIU/AEI/10.13039/501100011033/ y FEDER (UE).

F. X. Gaya-Morey contó con el apoyo de una beca FPU del Ministerio de Fondos Europeos, Universidad y Cultura del Gobierno de las Islas Baleares.

Referencias

Aechtner, J., Cabrera, L., Katwal, D., Onghena, P., Valenzuela, D. P., & Wilbik, A. (2022). Comparing User Perception of Explanations Developed with XAI Methods. *Proceedings of the IEEE International Conference on Fuzzy Systems – FUZZ-IEEE '22*, 1-7. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882743

Alqaraawi, A., Schuessler, M., Weiss, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. *Proceedings of the 25th International Conference on Intelligent User Interfaces – IUI '20*, 275-285. https://doi.org/10.1145/3377325.3377519

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J. T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2020). AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research*, *21*(130), 1-6. http://jmlr.org/papers/v21/19-1035.html

Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01276-x

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *Proceedings of the 38th International Conference on Machine Learning Research – PMLR '21*, *139*, 813-824. https://proceedings.mlr.press/v139/bertasius21a/bertasius21a-supp.pdf

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces – IUI '19*, 275-285. https://doi.org/10.1145/3301275.3302310

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '24*, 316.1-316.32. https://doi.org/10.1145/3613904.3642474

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI EA '22*. https://doi.org/10.1145/3491101.3503727

Gaya-Morey, F. X., Buades-Rubio, J. M., MacKenzie, I. S., & Manresa-Yee, C. (2024). *REVEX: A Unified Framework for Removal-Based Explainable Artificial Intelligence in Video*. https://doi.org/10.48550/arXiv.2401.11796

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(null), 1157-1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Heimerl, A., Weitz, K., Baur, T., & Andre, E. (2020). Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing*, *1*(1), 1-13. https://doi.org/10.1109/TAFFC.2020.3043603

Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim, J. (2020). ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS '20*, 10990-10997. https://doi.org/10.1109/IROS45743.2020.9341160

Kaplan, S., Uusitalo, H., & Lensu, L. (2024). A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems*, *283*, 111107. https://doi.org/10.1016/j.knosys.2023.111107

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset*. https://doi.org/10.48550/arXiv.1705.06950

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, *113*(523), 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '20*, 1-15. https://doi.org/10.1145/3313831.3376590

Liao, Q. V., & Varshney, K. R. (2022). *Human-centered explainable AI (XAI): From algorithms to user experiences*. https://doi.org/10.48550/arXiv.2110.10790

Liu, Z., Wang, L., Wu, W., Qian, C., & Lu, T. (2021). TAM: Temporal Adaptive Module for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision – ICCV '21*, 13688-13698. https://doi.org/10.1109/ICCV48922.2021.01345

Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences*, *12*(19). https://doi.org/10.3390/app12199423

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems – NIPE '17*, 4768-4777. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Manresa-Yee, C., Ramis, S., Gaya-Morey, F. X., & Buades, J. M. (2024). Impact of Explanations for Trustworthy and Transparent Artificial Intelligence. *Proceedings of the XXIII International Conference on Human Computer Interaction– Interacción '23*. https://doi.org/10.1145/3612783.3612798

Miller, T. (2019). Explanation in Artificial Intelligence: Insights From the Social Sciences. *Artificial Intelligence*, *267*(C), 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *11*, 24:1–24:45. https://doi.org/10.1145/3387166

OpenMMLab. (2020). *OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark*.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. *Proceedings of the British Machine Vision Conference – BMVC '18, Newcastle, UK*, 1-151. https://doi.org/10.48550/arXiv.1806.07421

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Ridley, M. (2025). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *76*(1), 98-120. https://doi.org/https://doi.org/10.1002/asi.24889

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *46*(04), 2104-2122. https://doi.org/10.1109/TPAMI.2023.3331846

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the International Conference on Computer Vision – ICCV '17*, 618-626. https://doi.org/10.1109/ICCV.2017.74

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. *Proceedings of the 26th International Conference on Intelligent User Interfaces – IUI '21*, 109-119. https://doi.org/10.1145/3397481.3450662

Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning: A systematic review of current approaches and trends. *Frontiers in Artificial Intelligance*, *4*, 1-15. https://doi.org/10.3389/frai.2021.550030

Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020). Temporal Pyramid Network for Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR '20*, 591-600. https://doi.org/10.1109/CVPR42600.2020.00067

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Proceedings of the 13th European Conference on Computer Vision – ECCV '14 (LNCS 8689)*, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53